Abstract

A video processing system tracks a moving person or other object of interest using a combined audio-video tracking system. The audio-video tracking system comprises an audio locator, a video locator, and a set of rules for determining the manner in which settings of a camera are adjusted based on outputs of the audio locator and video locator. The set of rules may be configured such that only the audio locator output is used to adjust the camera settings if the audio locator and video locator outputs are not sufficiently close and a confidence indicator generated by the audio locator is above a specified threshold. For example, in such a situation, the audio locator output alone may be used to direct the camera to a new speaker in a video conference. If the audio locator and video locator outputs are sufficiently close, the system determines if a confidence indicator generated by the video locator is above a specified level, and if so, the video locator output may be used to adjust the camera settings. For example, the camera may be zoomed in such that the face of a video conference participant is centered in and occupies a designated portion of a video frame generated by the camera.